LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
INSTITUT FÜR STATISTIK
SONDERFORSCHUNGSBEREICH 386

Pigeot, Heinicke, Caputo, Bruederl:

The professional career of sociologists: a graphical chain model reflecting early influences and associations

Sonderforschungsbereich 386, Paper 74 (1997)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

gsf

MAX-PLANCK-GESELLSCHAFT

TUM

THE PROFESSIONAL CAREER OF SOCIOLOGISTS: A GRAPHICAL CHAIN MODEL REFLECTING EARLY INFLUENCES AND ASSOCIATIONS

Iris Pigeot, Astrid Heinicke, Angelika Caputo, Josef Brüderl*

Currently, German universities take great interest in how their students perform in later professional life. Information on early determinants of success is needed in order to adjust educational programs and to cope more easily with the expected increase in the number of students in the coming years. In this paper we analyze data on the occupational careers of sociologists. The complexity of the underlying research question is taken into account by modelling the associations using so–called graphical chain models. These models are in general constructed such that conditional independencies can be concluded from the corresponding graph. We present the different steps in formulating the dependence structure. For checking its appropriateness, a model selection strategy is applied based on regression techniques. The final graph gives essential hints with respect to early determinants for professional success.

1. INTRODUCTION

Observational studies in the social sciences usually obtain a considerable number of variables for each individual under investigation. The analysis of dependencies and associations among the variables can be fairly easy in some situations, but in most cases it is very likely that the use of more sophisticated multivariate statistical

*University of Munich

Financial support by the SFB 386, Deutsche Forschungsgemeinschaft is gratefully acknowledged.

methods is required to capture the complex structure of the research question. The choice of an appropriate statistical model depends crucially on problem formulation and may be the most challenging phase of the statistical analysis.

As a matter of fact, observational studies are unable to reveal any causal relationship. This is of importance when it comes to interpreting the results of the statistical analysis with respect to the underlying substantive research hypothesis. This is a postulated scheme concerning the association structure of the variables motivated by subject–matter considerations of the investigator. If this research hypothesis can be confirmed one cannot conclude that this is due to an underlying causal relationship among the variables. Usually, there are also other structures among the variables that are in line with the data material. Nevertheless, a statistical analysis resulting in rejection of the research hypothesis makes clear that second thoughts are needed which may lead to modification and amendment of the hypothesis motivating further research. In that terms the causal structure is only based on a–priori reflections and subject matter knowledge available and is therefore independent of the data under analysis.

In observational studies, information is collected on many variables for each person. In general the variables of interest are divided into pure explanatory (influences) and pure response variables. Although this is the most common proceeding, we think that in our application there are good reasons to investigate also the associations among the explanatory variables. This is done not only by dividing the variables into pure responses which can be influenced by all remaining variables and into pure explanatories which are expected to have an impact on all remaining variables, but also into so–called intermediates which are simultaneously responses to some variables and influences to others. In recent years, graphical chain models which are able to cope with this complex type of situations have been introduced by the joint work of Cox, Lauritzen and Wermuth (Cox and Wermuth 1993, Lauritzen and Wermuth 1989).

Overall there are mainly three reasons for considering intermediates. The reasons are rather general arguments applying likewise to other studies. The point is that every investigator will have expectations concerning strength and direction as well as the absence of associations among the variables, no matter, if these are regarded as

responses or explanatories. That itself is an important motivation to take a closer look at relations among explanatory variables. The second reason for analyzing associations among the explanatories is to avoid wrong conclusions due to the well known Yule–Simpson Paradox (Simpson 1951, Wermuth 1987). This means that an association coinciding in several subgroups can be qualitatively different overall. The third and in our point of view one of the most striking arguments is that the introduction of intermediates makes it possible to find indirect influences. These are influences of variables that do not have an impact on the responses themselves but on other variables influencing the responses directly or perhaps also only indirectly. Such paths can be very helpful in getting to know early determinants of a certain behavior, attitude or other variables. Thus, overall consideration of intermediates enables the investigator to achieve a more detailed understanding of the situation.

Especially in our application the use of graphical chain models seems sensible. Our statistical analysis intends to find a detailed answer to the question: What makes sociologists be successful in their job? In order to find early determining factors we consider a large number of variables including biographical variables and variables concerning their professional career from the beginning of their studies up to now. As pure explanatories we have biographical variables like age and sex. The first group of intermediates consists of variables related to their time at the university like motivation, how long it took them to get their degree, whether they lived in a partnership. The variables covering the time between finishing university and the first job form the second group of intermediates: such as whether or not they have children or how much a successful professional life matters to them. Furthermore, we have the third group of intermediates consisting of variables related to their first job: whether they managed to find a job in the field they wanted to work in, or for how long they are employed. And finally, there are the variables of primary interest treated as pure responses: how satisfied they are with their job, their income and their sociological skills needed while doing their job.

A complete list of all variables can be found in Section 2, where also first descriptive results are presented. All these variables are assumed to have an ordering in time. The chronological order of the four groups is used for constructing the chain graph. The benefit of graphical chain models is twofold. In addition to the

ones already mentioned, they possess a graphical representation that is easy to interpret and provides a concise reflection of all conditional independencies proposed on a–priori arguments. This aspect is addressed in Section 3. The basic principles for constructing chain graphs are described in Section 3.1 without giving too many technical details. Section 3.2 deals with the derivation of the first rough association structure among the variables presented in Section 2. This structure is checked using a model selection strategy described in Section 4. The results of our analyses are given in Section 5. Interpretations and questions under current research are discussed in Section 6.

2. DESCRIPTION OF VARIABLES

We analyze data from the so–called "graduates study" being conducted at the University of Munich in 1995. In total 465 questionnaires were sent to sociology graduates having finished their studies between 1983–1994, of which 102 questionnaires could not be delivered, 89 of the remaining 363 were not answered which corresponds to a non–response rate of 24.52 %. We restrict on those questionnaires having no missing values in items being of interest for our analysis. Thus, we work with in total 182 complete cases.

In this section, we introduce the variables we consider as relevant to explain the current job situation of the graduates, where we focus on their scales and coding. Means, standard deviations, minimum and maximum values observed are shown in Tables 1, 2, and 3. In the tabular representation we distinguish the variables according to their measurement scales rather than dividing them into responses, intermediates or explanatory variables as we did earlier. The scale of a variable is of major importance since the appropriate choice of a statistical model depends crucially on it. Basically, variables are measured on a continuous or discrete scale. In our complex data set all kinds of variables occur: continuous, effectively continuous as well as ordinal, binary and polytomous ones.

We start off with the biographical variables. The variables *sex* (female = 1, male = 0) and *age* (continuous) of the student when entered university are expected to influence careers as well as some variables concerning the time period the now graduated persons spent at university. To find out if the graduates' motivation and

TABLE 1

Summary measures of the continuous variables.

variable	mean	standard deviation	minimum	maximum
earnings (DM/h)	34.80	11.55	6.16	73.13
adequacy	10.33	5.69	0	24.00
duration (years)	3.34	2.83	0	10.86
rank	19.08	4.77	7.00	29.00
assessment	22.41	3.46	9.00	32.00
experience (years)	2.45	2.61	0	9.14
uni	12.61	2.34	7.00	26.00
mark	2.10	0.50	1.00	3.30
computer	4.74	4.27	0	18.00
age (years)	22.58	3.57	19.00	36.00

attitude toward their studies do have an impact on later professional success, we ask whether their choice of taking up sociology was due to mere *interest* (1 = yes, 0 = no) toward the subject in general, or whether the decision was based on a special professional *aim* (1 = yes, 0 = no) they already had in mind. The graduates' *dedication* (1 = yes, 0 = no) to their studies is expected to give some hints about their attitude toward their field. We consider the time the graduates spent to get their degree (*uni*, continuous) as well as the *mark* (continuous) they achieved at their final exams and the *computer*-knowledge (continuous) they possessed by the time they left university. The variable *utility* (1 = yes, 0 = no) states the graduates' belief, if from a todays point of view the university played a dominant role for the development of certain non-sociological abilities like being able to work in a team, or to organize ones work efficiently, etc. Former analyses of the data have given some hints that the fact that graduates have a *partner*ship (1 = yes, 0 = no) at the time they left university has an impact on their later careers (Brüderl et al. 1996). The *branch* (polytomous) in which the graduates wanted to find a job, just after having completed their studies, is assumed to be influential. The variable has four

TABLE 2

Absolute and relative frequencies of the binary variables.

variable	0	1
match	88 (48.4%)	94 (51.6%)
child	156 (85.7%)	26 (14.3%)
interest	44 (24.2%)	138 (75.8%)
aim	154 (84.6%)	28 (15.4%)
dedication	77 (42.3%)	105 (57.7%)
utility	98 (53.8%)	84 (46.2%)
partner	82 (45.1%)	100 (54.9%)
sex	81 (44.5%)	101 (55.5%)

categories coded as three dummies. We distinguish between the area "university or research" which we use as reference and "industry" ($branch1 = 1$). Furthermore, we allow a category for graduates who consider both areas ($branch2 = 1$) as well as a category for graduates not having any preferences at all ($branch3 = 1$).

The next group of variables we introduce covers the time between the graduates left university and their current job. We regard as likely that having a *child* at the end of university (having at least one child = 1, zero elsewise) will make a difference to later job success. Also the graduates' attitude toward the importance of work on the one side and family life on the other is taken into account. It is measured by *rank* (continuous) where the higher the value the more work matters to them. The graduates were asked to assess a few not-university-related abilities they gained such as being able to work in a team etc. (compare also with variable *utility*). How much professional *experience* (continuous) the graduates gained before they have taken their current job is also assumed to have an effect on *earnings* (continuous).

The next group of variables consists of *match* (1 = yes, 0 = no), which equals one if a graduate settled down in the field he wanted to and zero otherwise. *Duration* (continuous) states for how long the graduates work at their current job and *contract* refers to the type of contract they have. The latter has three categories coded

TABLE 3

Absolute and relative frequencies of the polytomous variables. Categories are given below each variable.

variable *categories*	*frequencies*			
satisfaction	9	26	109	38
1, 2, 3, 4	(4.9%)	(14.3%)	(59.9%)	(20.9%)
contract	106	50	26	
1, 2, 3	(58.2%)	(27.5%)	(14.3%)	
branch	65	37	58	22
0, 1, 2, 3	(35.7%)	(20.3%)	(31.9%)	(12.1%)

as two dummies. We distinguish between a permanent employment (reference), a temporary employment ($contract2 = 1$) or if someone has started an own business ($contract3 = 1$). We operationalize success due to its multidimensionality with three variables. *Earnings* (continuous) represents the monetary aspect of success, whereas *adequacy* (continuous) measures how often graduates relate on their sociological skills in every day work. An overall measure of *satisfaction* (ordinal) refers to their current job situation and has four levels ranging from 1 to 4, where the lowest level corresponds to very dissatisfied graduates.

3. CONSTRUCTION OF THE CHAIN

3.1 Basic principles for deriving a graphical chain model

This section deals with the basic concepts for constructing a chain graph according to the rules given by Wermuth and Cox (1992). Such chain graphs are graphical representations of a statistical model reflecting under certain assumptions conditional or marginal dependencies among the involved variables, where we focus on conditional dependence graphs. In chain graphs, variables are represented by points and associations between each two of them as undirected lines or arrows. In the following we emphasize the rules for constructing such graphs and briefly point out relationships between chain graphs and their corresponding statistical models.

First all relevant variables have to be fixed. These variables are then as already mentioned represented by points, also called nodes in a graph theoretical context. Generally we distinguish between **c**ontinuous and **d**iscrete variables, the former are denoted by **c**ircles and the latter by **d**ots in the graph.

The groups of response, explanatory and intermediate variables are referred to as "chain links" represented by boxes. In the graphical representation, variables of one group, that are variables looked at as being on equal footing, own the same box. The boxes are arranged in a row according to the underlying rough association structure. Conventionally, the left–hand box contains pure responses, the right–hand box pure explanatory variables and in between these two extremes the intermediates are placed, one box for each level. The boxes of the different levels of the intermediates are ordered so that intermediates of a particular box are taken as potentially explanatory to all variables to their left as well as to be potentially responses to all variables to their right. Or to put it differently, variables of one box are considered conditionally on all variables of boxes to their right.

So far we represented the variables by points, placed them in boxes, arranged the boxes according to the underlying association structure resting upon subject matter knowledge. This association structure, also called dependence chain, determines the kind of associations possibly occurring among the variables. Variables of the same box can be only symmetrically associated. A symmetrical association between two variables belonging to the same box means in graphical terms, that these two variables are connected by a solid line without arrow–heads, also referred to as undirected edges in graph theory. Two variables belonging to different boxes can only be asymmetrically associated. This is represented by solid arrows with the arrow pointing from the influencing variable to its response. In graph theory, arrows are called directed edges. Note again that due to the underlying dependence chain all arrows point from right to left.

In general, directed or undirected edges mean that every particular variable of a box is investigated conditionally on all variables to its right as well as on all remaining variables of its own box. Two variables of the same box which are not linked are considered under certain model assumptions to be conditionally independent given all remaining variables to the right and of the same box. This precise meaning of

a missing link is ensured by the Markovian properties of concentration graphs, a special kind of graphical chain model we are interested in, as well as of the assumed distribution of the variables and usually does not hold for so–called LISREL–models (see also Chapter 6).

It should be mentioned, that a variable pair can be connected by one edge at most. And additionally, a chain graph does not allow for circles, i.e. a variable cannot be connected with itself which means that a variable cannot serve as explanatory to itself.

For more details about graphical chain models we refer for instance to Wermuth and Cox (1992), Lauritzen and Wermuth (1989), Cox and Wermuth (1996) or Lauritzen (1996).

3.2 The postulated dependence chain of the graduates–study

We now establish the rough association structure of the variables, that is we split them into responses, intermediates of different footing and explanatory variables according to our sociological assumptions. Since all further statistical analyses are based on this first structure, great care is needed concerning the choice of relevant variables (cf. Section 2). Due to their chronological order it was quite easy to formulate the first rough dependence chain as in Figure 1.

According to the rules mentioned earlier the biographical variables *age* and *sex* are placed within the same box at the very right since they are pure explanatories. They mark the earliest point in time of all variables and thus are unable to get influenced by any other variable of the model.

The first group of intermediates consists of variables covering the time the graduates spent at university. In our point of view, how one studies may have an impact on later career. *Dedication, interest* as well as *aim* relate to this aspect. Equally, achievement of qualifications beyond university courses, like *computer*–knowledge, appears to be relevant in this context. Further, we take interest in how much university contributes to development of skills useful and more or less indispensable for later professional success, which is captured by the variable *utility*. Other arguments apply to the variable *partner*. We assume that there is a mutual influence between private life and university demands. For example, we suppose that family planning

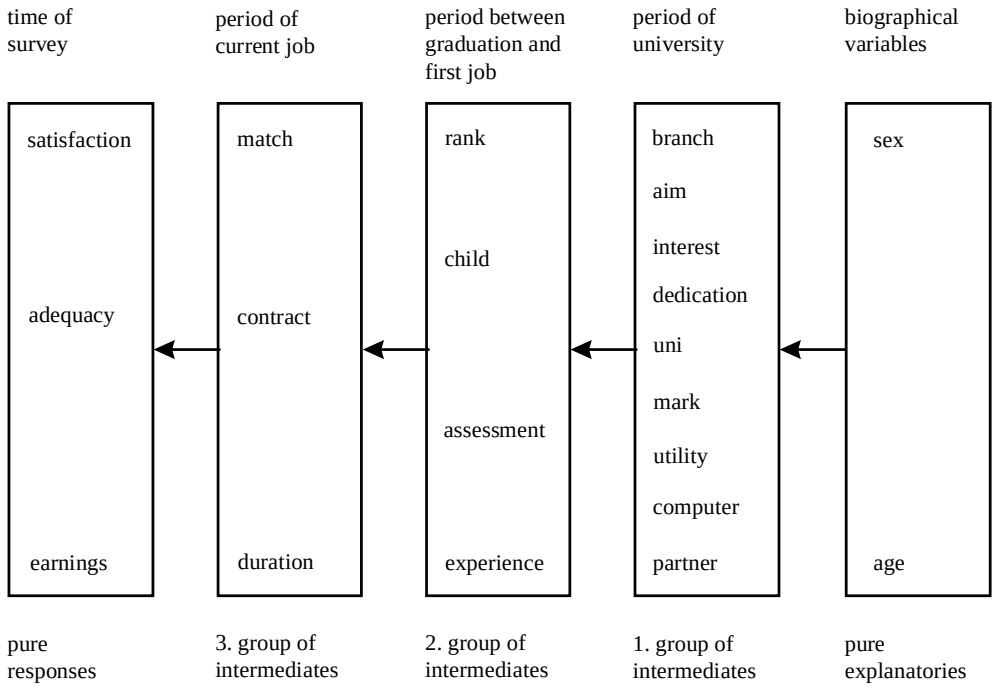

FIGURE 1 First rough association structure.

has an effect on how long it takes to get ones degree and on how long one is on job search later on, since private interests may override professional ones and might for example limit mobility, which itself is likely to cut down job chances. The variable *branch* is interesting for many reasons. It seems quite obvious, that working in a scientific area requires more sociological skills on a daily basis than elsewhere, but in general the payment is rather low compared to a job in the industry. Therefore, we are dealing with some kind of trade off between payment and sociological skills needed at work. Thus, a decision to work in a special branch means also a confession to certain preferences. We also expect some negative effects for graduates who do not yet know in which branch they want to work in. These students might not be as concentrated on planning their career compared to graduates having made up their mind who are expected to follow a certain strategy to achieve their goal.

The second group of intermediates covers the time between the graduates' first job and the point in time they have graduated. We think how a graduate values work

in general as well as having children point toward the amount of energy the graduate is willing to spend for the career which then is expected to affect professional success. We expect work experience achieved so far has a strong impact on our pure responses, especially on *earnings*. Further, we suspect graduates having high scores of the variable *assessment* are more likely to be successful.

The third group of intermediates consists of variables dealing with the current job. To see if a graduate managed to find a job in the branch she/he wanted to, we construct the variable *match*. How long someone already works at a particular place will certainly affect earnings, since with increasing duration experience and competence will grow. Probably, this will also lead to higher satisfaction and a higher requirement of sociological skills.

The responses are placed to the very left, because they can be influenced by all other variables directly or indirectly.

4. STATISTICAL ANALYSIS

4.1 Strategy

In the foregoing section the postulated dependence chain has been developed based on subject matter considerations. All implied direct, indirect, symmetrical, and asymmetrical dependencies require checking using appropriate statistical methods. Necessarily we proceed stepwise according to the dependence chain, thus five major steps have to be taken. We explore

- associations among the pure responses *satisfaction, adequacy*, and *earnings* as well as their dependence on all other variables,
- associations among the third group of variables *match, contract*, and *duration* as well as their dependence on all other variables to their right,
- associations among the second group of intermediates *rank, child, assessment*, and *experience* as well as their dependence on all other variables to their right,
- associations among the first group of intermediates *branch, aim, interest, dedication, uni, mark, utility, computer*, and *partner* as well as their dependence on all other variables to their right, and finally
- associations among the pure explanatory variables *sex* and *age*.

To derive a graph, block–recursive regressions (Wermuth 1992) have to be fitted. Block–recursive regressions enable us to analyze the dependencies among several responses and several influences using a system of univariate regressions only. These univariate regressions, which can be of all kinds depending on the scale of their particular response, describe altogether the association structure of the involved variables. The univariate regressions have in common that each of them has one response only and the explanatory variables consist of all variables to the right of the response as well as of the remaining variables of its own box. By using univariate regressions our analysis is close to the data, but the involved estimation procedure does not ensure the validity of the Markovian properties for the whole graph. Missing edges can therefore only be interpreted as conditional independencies for special cases. Thus, the resulting graph should be looked at from a more exploratory point of view. However, at the moment no better observation–driven strategy for analyzing this kind of data is known.

In the first step three univariate regressions have to be fitted, each of them considers one particular response at a time and all remaining variables as explanatories. To make this clear, say *earnings* serves as response, then all other variables are explanatory including *satisfaction* and *adequacy* as well. The three regressions as a whole describe the structure among the variables. In the second step there are again three univariate regressions and so on.

4.2 Variable selection strategy for one univariate regression

A good variable selection strategy leads to a model which has as few variables as possible and as much variables as necessary to describe the structure of the data in an appropriate way. On the one hand, a low number of parameter estimates means low noise for the number of possible estimation errors decreases and the costs for estimation of parameters not needed are avoided. On the other hand, the reduction of complexity is limited since a correct specification of the model, i.e. no neglection of essential variables, is desirable. Therefore, a selection strategy should provide some sort of an optimal solution to this trade–off. In general, automated selection strategies implemented in most statistical packages available have in common that only main effects which measure the influence of the explanatory variables on the

related responses are included or excluded. But, of course, variables may interact with each other with respect to their influence on the response. Thus, we think it has to be checked if interaction terms (i.e. mean cross product terms) have to be introduced into the model. Another problem which has to be accounted for in the modelling process concerns possible nonlinearities (here squared terms). In our case we used a selection strategy invented by Streit (1997) which we slightly modified. It combines common forward and backward selection strategies. The former strategy starts off with a "zero" model including one variable after another until no further variable fulfills a certain input criterion. The input criterion ensures that variables that do not increase the prediction of the response variable reasonably well are not taken in the regression equation. The philosophy of the backward strategy is to begin with a full model, i.e. a regression which includes all variables. Variables are then step by step excluded until no further variable fulfills some kind of exclusion criterion. Usually backward and forward procedures are carried out automatically and the criteria are based on an the change of r^2, the squared correlation coefficient which is used to measure the model fit, or on the increase of the likelihood depending on the kind of regression. This means that the choice between two models which might have a quite similar fit is typically not motivated by substantial considerations although it would be better to combine statistical strategies and subject matter reflections during the selection process. We favor the latter approach.

As already mentioned, we consider the additional check for nonlinearities and interactions to be essential. Therefore, use of a forward strategy means in our case that after the inclusion of all important main effects the check for nonlinearities and interactions will be based on variables already taken in. For that reason nonlinearities and interactions can be overlooked since they may depend on not inserted variables. The use of backward selection strategy is simply not practicable since one has to start off with a full saturated model and the number of parameters to be estimated is far too large in our case. Therefore, we proceed as shown in Figure 2. To get a hint which nonlinearities and interactions may need consideration we carry out two different screening–tests (Cox and Wermuth 1994). The idea of the screening is very similar to normal probability plots. For cross product terms we examine the t–values from trivariate regressions such as regressing Y on X_i, X_j and $X_i \times X_j$.

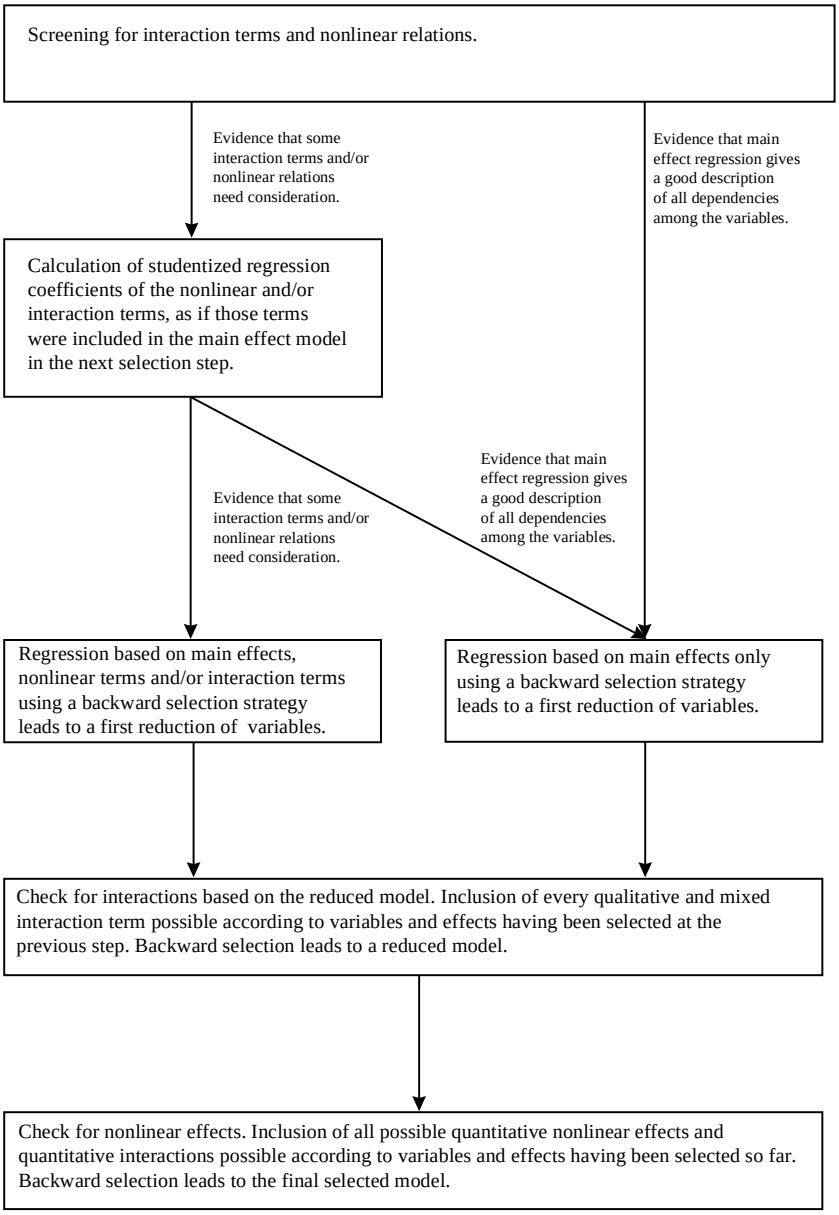

FIGURE 2 Schemata of the selection strategy applied to each univariate regression.

Due to the number of cases the t–statistics follow a standard normal distribution if there are no interactions. We then plot the ordered t-statistics against the expected values of the standard normal distribution. If the assumption is fulfilled, that there

are no interactions the points spread along the diagonal, whereas strong divergencies imply that there are interactions. We proceed likewise in checking for nonlinearities.

If the screening does not show a striking nonlinearity and/or interaction we start with a main effect regression. To be more precise we use a backward selection strategy. In each step we exclude the variable which has the smallest corresponding absolute t–value. We stop excluding variables as soon as all variables in the regression have a $|t|$–value greater than 2. Then, we check again for interactions and nonlinearities. First we include interactions involving qualitative variables as well as mixed interaction terms. This is followed by a backward selection as described above. Finally, we include all nonlinearities and all quantitative interactions possible and carry out a backward selection leading to the final reduced regression.

If the screening gives evidence that certain interactions or nonlinearities should be considered, we calculate their t–values, as if these terms were inserted in the main effect model in the next selection step. If the t–value is absolutely greater than 4 we include the corresponding interaction or nonlinearity in the next step and proceed with a backward selection as described above.

5. RESULTS

The screening shows a few interaction terms and nonlinearities that need further consideration. All estimated regression parameters are summarized in Table 4. We begin with the three response variables: *adequacy*, *earnings*, and *satisfaction*. The variable *adequacy* is influenced by the variables *match*, *assessment*, *rank*, *aim*, *branch*, and *satisfaction* as the univariate linear regression shows. We also fit a univariate linear regression to find the influencing factors to *earnings*. The variables *duration*, *experience*, *partner*, *sex*, *branch*, *contract*, *adequacy* as well as the interaction *adequacy* $\times$ *contract* prove to be influential. Since *satisfaction* is an ordinal polytomous variable a cumulative logit–model (McCullagh and Nelder, p. 102 1983) has to be used. The selection reduces the number of possible influencing variables to *adequacy*, *age* and *sex*. These regressions describe the direct dependencies and associations among the three pure responses and the rest of the variables altogether. They also imply the two following equivalent partial graphical representations (Figure 3). For subject matter reasons we favor representation a), because the adequacy

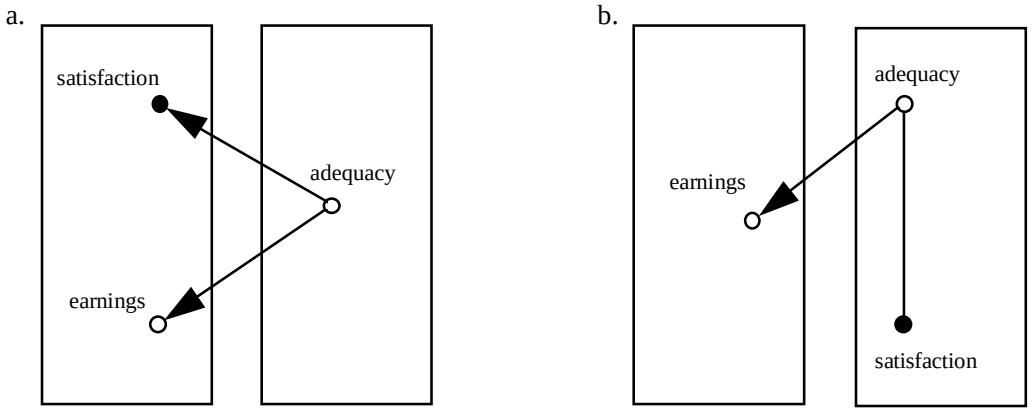

FIGURE 3 Two equivalent patterns of associations among the three pure response variables.

of a job is rather earlier in time than professional satisfaction. From a psychological point of view, one can argue that professional satisfaction influences the perception about how adequate the job is, but nevertheless this conclusion overstresses the data material at hand. The dependencies and associations obtained from the first step are shown in Figure 4.

We now examine the associations among the third group of intermediates, which consists of the variables *match, contract*, and *duration*. The linear regression having *duration* as response shows that the variables *duration, computer*, and *contract* have an impact.

Since the variable *match* is binary a logistic regression is fitted (Fahrmeir and Tutz, Chapter 2 1994) and the selection process shows *duration, computer, sex, branch*, and *contract* to be relevant. For the polytomous variable *contract* a multivariate logit–model is fitted. The variables *match, duration, mark, branch, rank, dedication, sex* as well as *uni* are influential. The graphical chain model corresponding to the last three regressions is represented in Figure 5.

The results for the second group of intermediates: *rank, child, assessment*, and *experience* are as follows. Amazingly *partner* is the only selected variable which has an impact on *assessment*. *Dedication, sex*, and *child* influence the variable *rank*, and

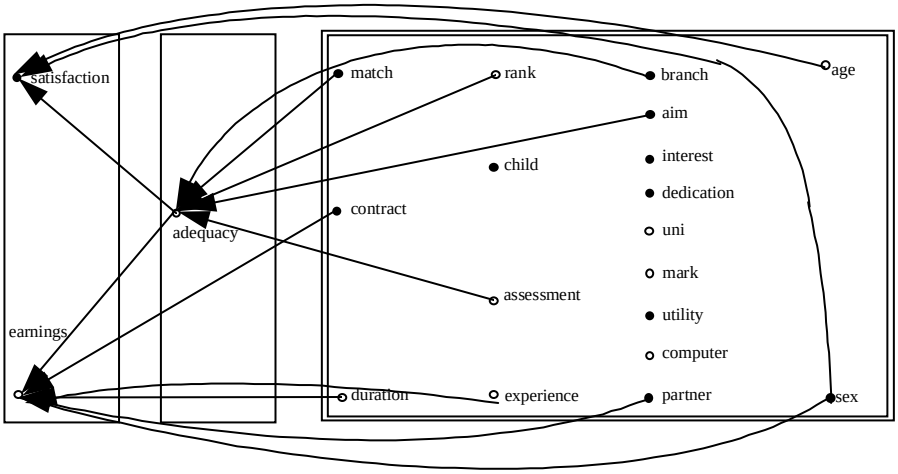

FIGURE 4 Graphical chain model of the three pure responses. The double box means that so far no statements about the associations among the included variables have been made.

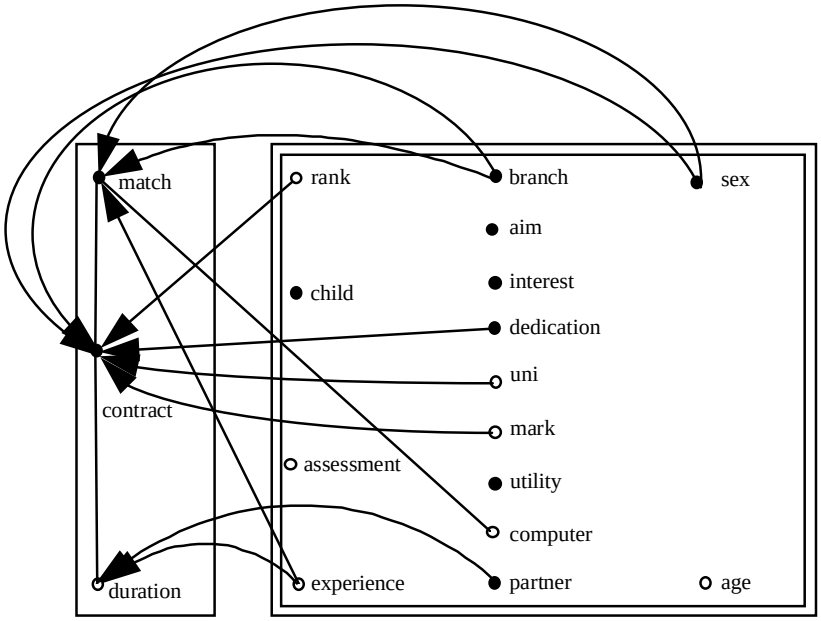

FIGURE 5 Graphical chain model of the third group of intermediates.

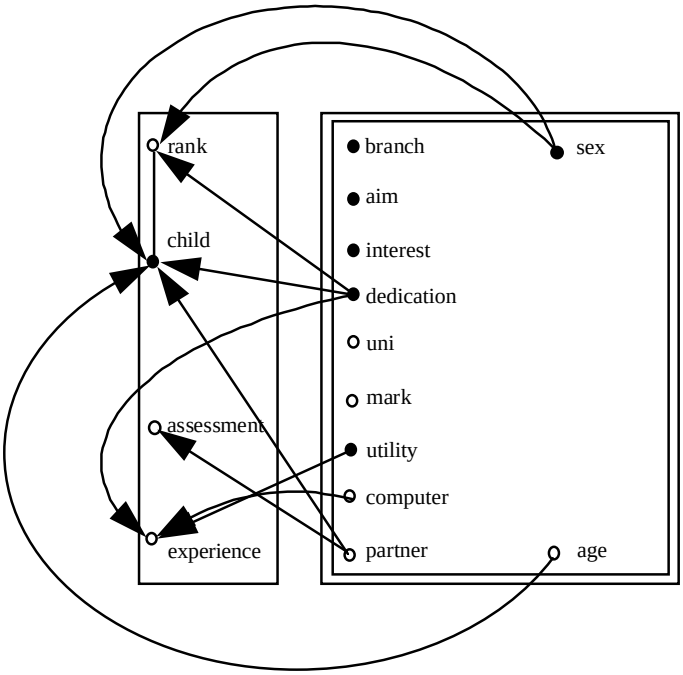

FIGURE 6 Graphical chain model representing the associations among the second group of intermediates and the variables to their right.

assessment, *utility*, *computer* affect the variable *duration*. For these three responses we fit one linear regression each. The binary variable *child* is modelled with a logistic regression. *Assessment*, *partner*, *age*, *sex*, and *rank* prove to be relevant. The graphical chain model corresponding to the second group of intermediates is shown in Figure 6.

The result of the first group of intermediates are now described. To each binary variable a logistic regression is fitted, to the polytomous variable *branch* a multivariate logistic model. The variables *mark*, *computer*, and *partner* are statistical important for the response variable *branch*. *Interest* is influenced by *aim*, *dedication*, and *utility* and *dedication* by *interest* and *mark*. The variables *dedication*, *age*, and *branch* have a considerable influence on *mark*. The variables *interest* and *computer* influence *utility*. The *computer*–knowledge can be explained by *utility* and *branch*. *Partner* is affected by *aim*, *age*, and *branch*. The graphical representation of these associations is shown in Figure 7.

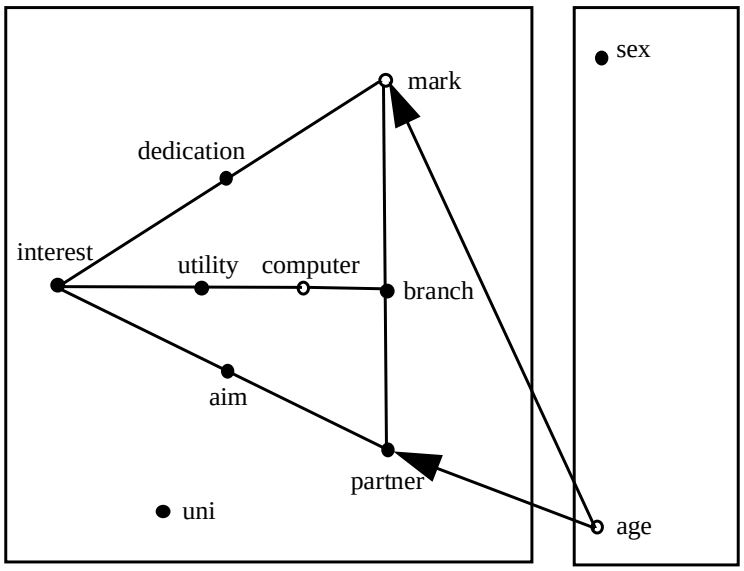

FIGURE 7 Graphical chain model of the associations corresponding to the first group of intermediates.

Between the pure explanatories *sex* and *age* at entering university no association is found.

TABLE 4:

Estimated regression coefficients $\hat{\beta}$ and

according standard errors $\hat{\sigma}_{\hat{\beta}}$ of the selected models.

response	*explanatories*						
adequacy	*match*	*assessment*	*rank*	*aim*	*branch0*	*branch2*	*branch3*
$\hat{\beta}$	2.83	0.31	0.19	3.17	–4.77	–2.50	–2.35
$\hat{\sigma}_{\hat{\beta}}$	0.77	0.10	0.08	0.98	0.98	1.01	3.28
	satisfact2	*satisfact3*	*satisfact4*	const			
$\hat{\beta}$	1.16	4.24	4.57	–9.69			
$\hat{\sigma}_{\hat{\beta}}$	1.84	1.64	1.77	3.28			
earnings	*duration*	*experience*	*partner*	*sex*	*branch0*	*branch2*	*branch3*
$\hat{\beta}$	1.77	1.85	2.85	–3.56	5.98	0.04	1.57
$\hat{\sigma}_{\hat{\beta}}$	0.24	0.26	1.32	1.27	1.90	1.80	2.34
	contract2	*contract3*	*adequacy*	*interact2*	*interact3*	*const*	
$\hat{\beta}$	–0.93	10.62	0.40	–0.13	–1.68	20.05	
$\hat{\sigma}_{\hat{\beta}}$	3.63	3.63	0.16	0.28	0.30	2.76	
satisfact	*adequacy*	*age*	*sex*	*const1*	*const2*	*const3*	
$\hat{\beta}$	–0.12	0.10	0.92	–5.01	–3.31	–0.12	
$\hat{\sigma}_{\hat{\beta}}$	0.03	0.04	0.31	1.11	1.05	1.02	
duration	*experience*	*computer*	*contract2*	*contract3*	*const*		
$\hat{\beta}$	–0.43	–0.19	–1.33	0.06	5.56		
$\hat{\sigma}_{\hat{\beta}}$	0.08	0.05	0.44	0.55	0.39		
match	*experience*	*computer*	*sex*	*branch0*	*branch2*	*branch3*	*contract2*
$\hat{\beta}$	–0.15	0.12	–0.98	1.01	1.19	–9.26	1.76
$\hat{\sigma}_{\hat{\beta}}$	0.07	0.05	0.38	0.55	0.55	19.40	0.51
	contract3	*const*					
$\hat{\beta}$	1.01	–0.70					
$\hat{\sigma}_{\hat{\beta}}$	0.53	0.67					
contract2	*match*	*duration*	*mark*	*branch0*	*branch2*	*branch3*	*const*
$\hat{\beta}$	2.47	–0.29	–1.17	–1.77	–0.17	1.39	2.47
$\hat{\sigma}_{\hat{\beta}}$	1.06	0.08	0.50	0.43	0.33	0.49	1.06
contract3	*rank*	*dedicat*	*sex*	*uni*	*const*		
$\hat{\beta}$	0.13	–0.58	0.47	0.22	–7.07		
$\hat{\sigma}_{\hat{\beta}}$	0.06	0.27	0.24	0.09	1.78		

Continued: TABLE 4

response	*explanatories*					
assessment	*partner*	*const*				
$\hat{\beta}$	1.08	21.82				
$\hat{\sigma}_{\hat{\beta}}$	0.51	0.38				
rank	*sex*	*dedicat*	*child*	*const*		
$\hat{\beta}$	1.72	1.52	1.74	17.25		
$\hat{\sigma}_{\hat{\beta}}$	0.70	0.71	1.00	0.68		
experience	*dedicat*	*utility*	*computer*	*const*		
$\hat{\beta}$	0.82	−1.33	−0.21	3.59		
$\hat{\sigma}_{\hat{\beta}}$	0.37	0.36	0.04	0.38		
branch0	*const*	*mark*	*computer*			
$\hat{\beta}$	−3.94	2.89	−0.25			
$\hat{\sigma}_{\hat{\beta}}$	1.21	0.61	0.06			
branch2	*const*	*mark*	*computer*	*partner*		
$\hat{\beta}$	−1.19	1.45	−0.18	0.35		
$\hat{\sigma}_{\hat{\beta}}$	1.09	0.56	0.05	0.17		
branch3	*const.*	*mark*	*computer*	*partner*		
$\hat{\beta}$	−3.36	2.12	−0.25	0.69		
$\hat{\sigma}_{\hat{\beta}}$	1.43	0.70	0.07	0.25		
interest	*aim*	*dedication*	*utility*	*const*		
$\hat{\beta}$	−1.09	1.29	1.07	0.27		
$\hat{\sigma}_{\hat{\beta}}$	0.48	0.38	0.39	0.29		
aim	*interest*	*partner*	*const*			
$\hat{\beta}$	−1.01	1.16	−1.73			
$\hat{\sigma}_{\hat{\beta}}$	0.45	0.48	0.47			
dedicat	*interest*	*mark*	*const*			
$\hat{\beta}$	1.13	−1.87	3.45			
$\hat{\sigma}_{\hat{\beta}}$	0.39	0.39	0.88			
mark	*dedicat*	*age*	*branch0*	*branch2*	*branch3*	*const*
$\hat{\beta}$	−0.32	0.03	0.42	0.16	0.26	1.46
$\hat{\sigma}_{\hat{\beta}}$	0.07	0.01	0.09	0.09	0.12	0.21
utility	*interest*	*computer*	*const*			
$\hat{\beta}$	1.06	−0.09	−0.58			
$\hat{\sigma}_{\hat{\beta}}$	0.38	0.04	0.38			
computer	*utility*	*branch0*	*branch2*	*branch3*	*const*	
$\hat{\beta}$	−1.13	−4.15	−3.11	−4.13	−1.13	
$\hat{\sigma}_{\hat{\beta}}$	0.59	0.82	0.84	1.06	0.59	
partner	*aim*	*age*	*branch0*	*branch2*	*branch3*	*const*
$\hat{\beta}$	0.99	0.11	0.31	−0.49	−1.10	−2.19
$\hat{\sigma}_{\hat{\beta}}$	0.48	0.05	0.44	0.44	0.58	1.13

6. DISCUSSION AND CONCLUSION

Summarizing our results, we think, that the use of graphical chain models has given way to deeper insights into the complex structure of the research problem. As an example, we discuss the following path out of the whole graphical chain. We found that sociologists seem to belong to one of two groups. One group consists of students who took up sociology having a certain professional goal in mind, whereas the other started to study sociology out of interest on the subject per se. The latter having higher scores of *dedication* and *interest* consider the university to be rather useful in terms of how it supplied them with extra non–sociological, but essential skills. Nonetheless, they tend to have lower developed computerskills and to a certain degree they lack of having a certain idea of what job to take in later professional life. Persons who already know what kind of job they want to take have higher scores of *adequacy*.

Overall the detailed analysis using graphical chain model suggests early determining factors of success in professional life. Analyses with the aid of classical multivariate analysis techniques would not have been able to show these early factors since they cannot model indirect influences, or cannot be applied in situations like the one we were confronted with. The latter holds also for LISREL–models, since they do not allow for modelling binary response models like logistic regressions, for example. Nevertheless, we want to compare briefly these two approaches which are in some situations competing and in others equivalent (Cox and Wermuth 1993). The possibility to model loops is one advantage of LISREL, for it supplies greater options in formulating research questions. Until now, for graphical chain models no comparable theory has been developed. One often mentioned argument in favor of LISREL is that global tests exist to check the model fit. We do not think that this is a real advantage of LISREL–models, because such a test does not reveal in which parts a LISREL–model does not describe the data well. For the selection strategy, however, this is of importance, because applying a graphical chain model scientists are forced to have a very precise idea of all possible associations that are likely to occur among the variables. This leads to a very structured selection process. But since the global tests connected to LISREL do not tell in which part the

model fit is not satisfactory, variable selection within LISREL often occurs rather intuitively and is more guided by trial–and–error. As mentioned above the interpretation of the parameters differs for both models leading to a different interpretation of the graphical representations. Using LISREL–models missing links cannot be interpreted as conditional independence of the related variables. For an interesting example concerning this problem see van de Geer (1971).

It is still an open question how to handle missing values in the context of graphical chain modelling. We conducted a complete case analysis and due to the large number of variables the decrease in cases was rather high. Thus, finding a way how to deal with missing values is important especially if a high number of variables is involved.

Another challenging problem concerns the modelling of event history data which would have been also of interest for the questionnaire investigated here. Because of the lack of an adequate statistical theory related to the graphical models for event history data, we neglected information on the transitions in the professional careers of the graduates.

REFERENCES

Brüderl, Josef, Thomas Hinz, and Monika Jungbauer–Gans. 1996. "Langfristig erfolgreich. Münchner Soziologinnen und Soziologen auf dem Arbeitsmarkt." Soziologie 3:5–23.

Cox, David R., and Nanny Wermuth. 1993. "Linear Dependencies Represented by Chain Graphs." *Statistical Science* 8:204–283.

——-. 1994. "Tests of Linearity, Multivariate Normality and the Adequacy of Linear Scores." *Applied Statistics* 43:347–355.

——-. 1996. *Multivariate Dependencies: Models, Analysis and Interpretation.* London: Chapman and Hall.

Fahrmeir, Ludwig, and Gerhard Tutz. 1994. *Multivariate Statistical Modelling Based on Generalized Linear Models.* Berlin: Springer.

Lauritzen, Steffen L. 1996. *Graphical Models.* Oxford: Oxford University Press.

Lauritzen, Steffen L. and Nanny Wermuth. 1989. "Graphical Models for Associations Between Variables, Some of Which Are Qualitative and Some Quantitative." *The Annals of Statistics* 17:31–57.

McCullagh, Peter, and John A. Nelder. 1983. *Generalized Linear Models.* London: Chapman and Hall.

Simpson, E. H. 1951. "The Interpretation of Interactions in Contingency Tables." *Journal of the Royal Statistical Society*, Series B, 13:238–241.

Streit, Reinhold. 1997. *Graphische Kettenmodelle mit binären Zielgrößen: Modellierung und Datenbeispiele in psychologischer Forschung.* Lengerich: Pabst. To appear.

Van de Geer, J. 1971. *Introduction of Multivariate Analysis for Social Sciences.* San Fransisco: Freeman.

Wermuth, Nanny. 1987. "Parametric Collapsibility and the Lack of Moderating Effects in Contingency Tables with Dichotomous Response Variable." *Journal of the Royal Statistical Society*, Series B, 49:353–364.

——-. 1992. "On Block–Recursive Linear Regression Equations." *Revista Brasileira de Probabilidade e Estatistica* 6:1–56.

Wermuth, Nanny, and David R. Cox. 1992. "Graphical Models for Dependencies and Associations." In *Computational Statistics*, edited by Y. Dodge and John Whittaker, 1:235–249. Heidelberg: Physica.